金融管理研究

王周伟 主编

第11辑 | VOL.11

FINANCE AND MANAGEMENT RESEARCH

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

金融管理研究

第 11 辑

王周伟　主编

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

内容提要

本书为研究金融管理的文集，包括12篇文章，内容包含货币政策、互联网金融、上市公司治理、人力资本、县域经济等论题，每篇文章的选题都是当前中国经济的热点，对读者有借鉴意义。

本书适合金融专业研究人员参考阅读。

图书在版编目(CIP)数据

金融管理研究.第11辑 / 王周伟主编. —上海：上海交通大学出版社，2018

ISBN 978-7-313-19703-0

Ⅰ.①金… Ⅱ.①王… Ⅲ.①金融管理-研究 Ⅳ.①F830.2

中国版本图书馆CIP数据核字(2018)第153654号

金融管理研究（第11辑）

主　　编：王周伟
出版发行：上海交通大学出版社　　地　　址：上海市番禺路951号
邮政编码：200030　　电　　话：021-64071208
出 版 人：谈　毅
印　　刷：上海春秋印刷厂　　经　　销：全国新华书店
开　　本：710mm×1000mm　1/16　　印　　张：18
字　　数：267千字
版　　次：2018年8月第1版　　印　　次：2018年8月第1次印刷
书　　号：ISBN 978-7-313-19703-0/F
定　　价：69.00元

本丛书是上海师范大学商学院组织的经济、金融与管理领域的理论研究与实践创新领域的中英文专业系列学术丛书，主要收录经济、金融与管理领域内原创性强、高质量的规范分析与实证研究方面的学术论文，也包括案例分析、文献综述、调研报告等。

主办单位：上海师范大学商学院

通信地址：上海市桂林路100号上海师范大学商学院

邮政编码：200234

联系电话：021-64321709

网　　站：www.jrglyj.com

邮箱地址：jrglyj@shnu.edu.cn

编　委　会

前　言

《金融管理研究》(Finance and Management Research)是经济、金融与管理理论研究与实践创新领域的中英文系列专业学术丛书。由上海师范大学商学院主办，每年出版2辑，每辑20万字左右。本丛书秉持学术性、创新性与前瞻性，努力为经济与管理研究搭建一个具有国内外领先水平的学术交流平台。

《金融管理研究》主要收录经济、金融与管理领域内，原创性强、质量高的规范、实证的学术论文，也可以是案例分析、文献综述、调研报告等。

本书集合了大量探索性、提炼性和创造性的科研工作，希望对中国经济与管理学界的交流与发展能够有所裨益。

目　录

信用风险研究

基于决策树信用卡风险管理的研究　/李杨芝/ 1
基于 KNN 的公司流动性风险识别研究　/谢佳芳/ 22
基于 BP 神经网络的商业银行客户信用风险评估　/曹　培/ 46
基于支持向量机分类预测的上市公司债信用评级研究　/徐闪赏/ 63
基于 KMV 模型和关联规则组合的上市公司信用风险传染研究　/吕兵静/ 83
基于数据可视化 Rattle 的个人信用风险评价建模　/赵海鹏　李　丹/ 112

证券市场风险研究

开放式基金风险评级研究
——层次聚类法和随机森林算法的应用　/朱　青/ 137
上市公司经营风险甄别研究
——基于大数据机器学习　/赵晓媚/ 153
基于一般聚类分析的中国金融系统性风险评估研究　/汪　平/ 177
投资者舆情指数对股价波动风险影响的研究　/刘少伟/ 200
基于 TOPSIS 法与支持向量机的中国制造业股票价格波动风险识别　/刘　鑫/ 228
基于优化 SVM 算法的上证 180 指数选股策略构建　/陈　莹　梁　成/ 251

《金融管理研究》约稿启事　/275

基于决策树信用卡风险管理的研究[①]

李杨芝[②]

摘　要：自从信用卡引入我国以来，其发展速度与日俱增。伴随着我国经济的发展和国民消费习惯的改变，信用卡业务将成为人们生活中不可或缺的一部分。为了让信用卡更好地为经济发展和人们生活服务，信用风险管理就显得尤为重要。因为信用风险是信用卡业务的主要风险。所以发卡机构应在信用风险管理方面加强控制，建立一套合理有效、科学完善的信用评价体系。一般情况下，信用评价会将客户划分为“好”或者“坏”客户。根据客户的个人信息和历史数据，通过数学方法进行建模，进而来预测信用卡客户的违约风险。在这方面，决策树算法正确率高，而且直观简单。本文使用决策树作为建立信用评价模型的方法。本文首先对涉及的文献进行了整理归纳，将信用卡风险管理理论、决策树的理论框架进行结合，分析了决策树、信用卡业务中信用风险的特征。采用决策树 C4.5、C5.0、rpart、xgboost、组合预测模型算法进行建模，并且和支持向量机进行了比较。采用决策树方法构建的个人信用评价模型具有精度高、操控性强，并且在实践中使用广泛等诸多优点。

关键词：信用卡；决策树；风险管理

---

① 本文得到了王周伟教授、傅毅副教授的悉心指导，特此致谢！

② 李杨芝：1991.8－，男，硕士研究生，上海师范大学商学院。研究领域：金融发展、金融市场与金融机构。

1 引言

近年来，信用卡业务发展迅速。伴随着居民消费习惯的改变和人均收入的提高，信用卡将继续保持着增长势头。但是随之而来的信用风险也将成为发卡机构所面临的主要风险。近年来我国经济处于转型阶段，经济承受着下行压力，信用卡贷款的资产质量也慢慢承受着压力。居民的消费卡使用以及还款行为与公司层面的经营状况有很高的相关性。但银行发行信用卡数量巨大，极度审慎的资质审核成本过高，因此客户与银行双方存在信息不对称的博弈。往往银行都是在恶意透支及违约事件发生后才进行业务干预，但这事后的亡羊补牢无法追回全部损失，反而很大程度上会造成较大损失。因此，如何有效地对信用卡客户进行风险识别和检测，控制潜在的损失，银行需要在理论与实践两方面进行努力。

这样，一方面对于银行风险控制体系的职能有促进，对银行自身的持续经营与盈利能力有提高；另一方面，银行安全体系的逐步构建与完善使得银行自身更加适应快速发展的金融市场和经济全球化进程，能更好地向经济发达体的商业银行看齐，对于我国银行的全球发展战略有很大帮助。

国际市场中的信用卡发展历史较为悠久，发达金融体系中都有相对成熟的信用卡风险鉴别与检测的模型。但鉴于我国信用卡发展历史较短，国际通用的风险计量模型并不能适用于中国信用卡市场，因此我国银行业急需探索出适合我国国情的信用卡风险管理计量体系。

2 文献综述

2.1 关于信用卡风险管理的文献综述

Stiglitz 和 Weiss(1989)[1]研究发现信息不对称的因素导致银行与客户存在极端不平等的契约关系，客户融资行为带来的潜在收益没有上限约束，但损失存在确定下限，银行角度与此相反。这种不平等带来的期望收益使得客户“道德风险”骤升。银行一般的解决手段只能是提高借款利率，但这种方法的约束仍然有限，并且会损失优质客户资源，造成“逆向选择”。

Elizabeth Langwith(2005)[2]的研究发现影响信用卡业务风险高低的主

因为融资利率水平过高。银行通过降低持卡人承担的借款利率、平衡自身发展规模能够降低其所面临的信用风险。

Nadia Massoud 和 Anthony Saunders(2011)[3]的研究发现信用卡惩罚性费用能一定程度上反映银行市占率,并且该费用能体现并替代市场实际利率。

胡勇、张永青(2006)[4]的研究主要聚焦于个人消费信贷,并发现机构间风险信息共享、机构内管理水平提升、行业间征信系统建立以及提高金融科技化水平都能有效降低信用风险。对于已存在的过高风险资产,应通过资产证券化手段及时转移,降低银行经营风险。

闫天兵、沈丽(2008)[5]通过研究信用卡业务盈利方式提出信用评分机制模型,为银行量化风险提出指导。

王保艳(2010)[6]基于法律政策角度提出银行信用卡业务风险控制应从提高监管力度与效度、完善法律政策体系着手。

魏鹏(2011)[7]通过研究外国市场信用卡业务的风险特征结合中国市场现状,对行业内相互促进、行业间合作共赢以及银行内部管理机制等方面提出建议。

2.2 关于决策树研究的文献综述

决策树模型是最为常用的数据挖掘模型。以树状结构表现各属性变量的重要程度,从大量包含白色噪声的原始数据集中发现隐含的信息和规律。

Hunt.E.B 提出了决策树算法的基础—CLS,增加逻辑判定的结点数量来丰富原始空的决策树,这一过程一直进行直到所有训练集数据被正确分类。J.R.Quinlan 在 ID3 基础上提出 C4.5,该方法着重改进了缺失值处理、噪声数据适用性以及决策树剪枝规则。Leo Breiman 等在 ID3 基础上又提出 CART 算法,该方法由于能够应用在数据挖掘和预测中,因此对于复杂且量级大的数据具有很大的优越性。

近年来国内对数据挖掘的研究也日益深入,研究的主要方向在于各因素的权重划分和如何选择节点。黄定轩(2003)[8]认为在处理多因素问题研究中各因素的权重分配对数据挖掘的精度和效率有重要影响,因此他提出使用客观信息熵对多因素权重进行分配的方法,通过对大量的数据模糊聚类,运用粗糙集理论中的信息熵概念对实际数据中进行客观地评价,确定各因素的权重。张品

(2012)[9]对多重组合分类决策树算法进行了改进,他提出了基于遗传算法的方法,决策树的分类精度有显著的提高。姚亚夫、邢留涛(2011)[10]对连续数值的属性最佳分割阈值提出新的选择算法,基于连续变量最佳割点在边界的性质,选择相应变量最佳分割阈值,并且对C4.5分类其进行训练改进,在人车目标识别中有所应用。钱网伟(2012)[11]深入分析了决策树模型经典算法ID3生成算法的可并行性,他使用云计算的MapReduce编程技术实现了对海量数据的并行算法。张启徽(2015)[12]针对Apriori法内生的频繁项集弱点,提出相应修剪与优化方案,减少频繁项集,提高连接速度,通过增加项数统计字段和对不再使用的子项在数据库中标记或删除等方法使数据库数据规模不断减少,从而缩小搜索范围,提高扫描速度。

2.3 关于数据挖掘在信用卡风险管理应用中的文献综述

信用评分模型经过几代学者的探索和研究后在理论和技术上都趋于成熟。David Durand根据客户的信用卡贷款历史记录,评价信贷风险,实现了对"好"的贷款和"坏"的贷款的区分。20世纪80年代,随着数据挖掘理论的发展,越来越多的商业银行开始运用以决策树来建立信用风险水平评价体系。其中富国银行通过C5.0构建了客户属性与信用业务之间的关系体系;花旗银行使用CART算法对贷款客户的历史信贷数据进行分析,得出影响贷款的主要风险因素。

近年来,国内对如何应用数据挖掘技术到信贷风险的研究日趋增多。石庆众、靳云汇(2006)[13]实证研究结论得出决策分类树和神经网络等非传统线性方法在银行信用评分体系构建问题上有明显的精度优势。

徐晓霞、李金林(2006)[14]使用中国商业银行数据使用决策树C4.5建立银行风险评估模型,使用相应财务指标作为输入变量,进行违约与否的二项分类。

姜明辉(2007)[15]采用小样本数据比较了K-近邻法、神经网络法和分类树法,实证结果发现分类树算法在变量选择和分类精度上明显优于其他两种算法。

冯琼、叶涛(2009)[16]研究集中在客户价值管理与树挖掘的结合可能性。研究发现数据挖掘技术能够在效率和准确性两方面全面超越传统方法,不仅能够细分现有客户价值体系,降低成本,还能扩大客户群体,精准发现潜在客户,

创造新的业务盈利点。交叉研究还能展现出关联关系，为银行的业务营销提出新方向，最后还能通过机器学习总结出风险特征，并对新客户进行基于历史数据的筛选判别，为银行风险控制提出新方案。

庞素琳、巩吉璋(2012)[17]通过研究加入了 Boosting 技术的 C5.0 算法，成功构造成本矩阵和成本敏感决策树，并应用于商业银行个人授信体系。

3 模型的原理介绍

至今，决策树模型是使用最为广泛的归纳分类算法之一，该模型为了到达预测的目的会先对目标数据进行分类，依据训练集形成决策树。第一次形成的决策树并不一定能够对每个数据进行有效的分类，那么我们需要调整训练集，加入一些例外到训练集，通过重复此操作直到形成正确的分类为止。决策树模型还有一个很大的优点就是它能够处理字符等非数值数据，而其他一些模型需要先将非数值数据转换为数值型数据才能分析，决策树就省去了数据的预处理。此外，决策树给出的结果直观易懂，可读性强，很容易就能转成 SQL 语句。

当前，决策树发展突飞猛进，形成了诸多算法。如 CART、ID3、C4.5、C5.0、Xgboost 等。在这些算法中，当数 ID3 算法使用最为常见、最具代表。该算法是由 Quinlan J R 在 1986 年提出。随着决策树的进一步发展，其预测精确还将进一步提高，并且运用范围会更加广泛。

3.1 ID3 算法

信息论是 ID3 算法的理论基石，此算法是以信息熵和信息增益为依据来选择重要属性。根节点的选取是根据最大信息增益。再由不同的根节点对决策树进行分枝，最后对每个分枝采用相同的递归算法，以达到对数据归纳分类的目的。其思路如下：

假设集合 S 中含有 s 个样本，而每个样本只有两个属性：正和反，记为 L_i，$i=1$，2。具有 L_i 属性的样本共有 S_i，则对目标数据分类需要处理的信息量为：

$$I(S_1,S_2)=-\sum_{i=1}^{2}p_i\log(p_i) \tag{1}$$

假如属性 X 有$\{x_1,x_2,\cdots,x_n\}$个取值。我们根据取值的不同把 S 分为 n

个子集$\{S_1,S_2,\cdots,S_n\}$，S_j表示集合S中X属性是x_j的样本，如果X作为测试属性，令S_y是S_j中属于L_i的样本，则根据X分出的样本所需要的信息熵为：

$$E(X)=\sum_{i=1}^{n}\frac{S_{1j}+S_{2j}}{S}I(S_{1j},S_{2j})=-\sum_{j=1}^{n}\sum_{i=1}^{2}\frac{S_{1j}+S_{2j}}{S}p_{ij}\log(p_{ij}) \quad (2)$$

其中：$p_{ij}=\frac{S_{ij}}{|S_j|}$表示$S_j$中任意样本属于$L_i$的概率。信息熵越小表示数据越具有规则性，信息熵越大表示数据越杂乱无章。

由此规则得到的信息增益为：

$$Gain(X)=I(s_1,s_2)-E(X) \quad (3)$$

信息增益表示有效信息熵的损失量。该值越大表示目标属性根据X属性进行分类所损失的信息熵越大，这就意味着此属性越需要在决策树模型的上面。

ID3算法使用从上到下不回头的方法，穷尽可能的决策，这样能够确保得到一个简单有效的决策树。而能够保证每一步都能找到最佳属性的依据便是信息增益。把某一属性作为节点的标准为：选择最大信息增益的属性进行分类，通过逐一检验各个待检属性，采用同样的处理一层一层地构造决策子树$S_1,S_2,\cdots,S_n$。

虽然ID3算法存在很多优点，但是其也有着比较突出的劣势：

(1) 以信息增益的大小来选择属性会使得ID3更加偏向选择取值多的属性。

(2) ID3对于连续性数据无能为力，故在进行建模时需要将连续性数据离散化，这样就削弱了ID3算法的简单性。

3.2 C4.5算法

任何模型都会在实践过程中慢慢地改进。对于ID3算法的缺点，1993年的时候Quinlan J R就对其进行了改进，形成了C4.5算法，C4.5算法在此后风靡一时。

(1) 不同于ID3算法，C4.5把信息增益比作为衡量标准来挑选属性，这样便有效地消除了多值趋势，属性X所具有的信息熵为：

$$SplitInfo(X)=-\sum_{i=1}^{n}\frac{S_{1j}+S_{2j}}{S}\log_2\frac{S_{1j}+S_{2j}}{S} \tag{4}$$

那么属性 X 相对应的信息增益比为：

$$GainRatio(X)=\frac{Gain(X)}{SplitInfo(X)} \tag{5}$$

C4.5 采用信息增益比来选择属性进而得出决策树，该方法选取的任意节点都是信息增益比最大的属性。就是这看似简单的改进，使得 C4.5 算法比 ID3 更加简单有效，可靠性更强。

(2) 同时 C4.5 算法可以直接对非离散化的数据进行离散化处理。其操作为：对连续属性 A，C4.5 会找到一个最优闭值 T，在通过比较 A 与 T，构建 $A\leq T$(左支)和 $A\geqslant T$(右支)这两个分支，其中 T 称为分割点，就像用电锯将一块完整的木头一锯为二。这种方法不仅能够对连续数据离散化处理，并且找到了最优的分割点。该分割点是模型自动找寻，相对于人工找寻，更加便捷准确。

同样，C4.5 也并非一种完美的算法，其构建决策树的效率并不高，这样对于大数据就耗时较长。这是因为 C4.5 算法对数据集的扫描和排序是多次的。这种特性便决定了 C4.5 适用的数据集是能够驻留在内存的。

3.3 C5.0 算法

有缺点的地方，人们必定会想方设法进行完善，而 C5.0 便是在 C4.5 的基础上进行了优化，不仅大大提升了效率，而且适用于大数据建模。

C5.0 算法对 C4.5 算法进行了改进完善，使得其适合大数据。该算法运用了 boosting 方法，所以又被叫做 boostingtrees，boosting 方法能够提高模型的有效性和精度，运行速度快，占用内存小。

C5.0 算法具体决策树的构建思想如下：

(1) 以给定样本集合作为决策树根节点。

(2) 分别算出每个属性的信息增益率，把最大信息增益比值作为节点的分裂属性。

(3) 对上述属性的各个数值都建立一个分支，进而将样本分成 n 个子集，并且创建每个子集的节点。

C5.0 算法把信息熵的减少速度作为衡量标准来确定最优分支变量和分割

点值的依据。信息熵的减少反映了信息确定性增加不确定性下降，故停止分枝的时点是根据信息熵的减少速度。

C5.0算法是在此前算法的基础上，针对不足进行了优化改进，具有以下特点：

(1) C5.0算法处理缺失数据和字符数据等问题时稳定性更强。

(2) C5.0算法能够高速并行处理大量离散型和连续型数据。

(3) C5.0算法更加通俗易懂，上手较快，结果可读性强。

C5.0算法使用了boosting技术，使得其精度更高。

3.4 rpart算法

在R语言软件上并没有默认安装rpart的包，如需使用此算法，需要另行下载安装。rpart算法的函数并不多，主要的函数就两个。一个是rpart()函数；另一个是prone()函数。

rpart()函数具体表示为：rpart(formula，data，weights，subset，na，action=na，rpart，method，model=FALSE，x=FALSE，y=TRUE，parms，control，cost，...)；prone()函数具体表示为prune(tree，...) prune(tree，cp，...)，前者用来拟合，后者用来剪枝。剪枝的目的是防止过度拟合，从而避免生成的树完全拟合了原始数据，如果树的枝太多就会完成拟合了原始数据，就不能反映数据本身的规律性，不具有现实意义。如果枝太少，那么预测的精确将大打折扣。

rpart算法停止运行的依据是达到了给定条件，我们会事先设定约束，即偏差小于设定阈值、节点中的样本数小于设定阈值、决策树的深度大于设定阈值。这三个约束条件是根据rpart()函数的三个参数(cp、minsplit、maxdepth)确定，如果不设定，其默认值分别为0.01、20、30。但是默认值并非是最优，我们为了防止过度拟合常常会调整参数的数值。所以对构建的决策树进行事后修建是至关重要的。

我们选择决策树方法有二，一是选取交叉验证的相对方差(xerror)最小值；二是根据剪枝理论中运用相对广泛的1－SE(1标准差)规则，该规则是指：在交叉验证的预测误差尽可能小的情况下，选取尽可能小的复杂度参数，然后根据此复杂度参数进行剪枝。这里的预测误差尽可能小并不意味着最小，而是在一个标准差范围内尽可能小。然后在此范围内选择所需的复杂度。第二个

规则不仅能保证误差尽可能小，而且还考虑了复杂度参数尽可能小。之所以这样选取是因为随着拆分的增加，复杂度会减少，而预测误差会先减少后增加。故无法同时保证复杂度和预测误差都最小。因此我们在预测误差某个尽可能小的范围内选择尽可能小的复杂度。

3.5 xgboost 算法

xgboost 的全称是 extreme gradient boosting，损失函数表示为 $l(y_i, \hat{y}_i)$，函数复杂度表示为 $\Omega(f_k)$。Objective：$\sum_{i=1}^{n} l(y_i, \hat{y}_i) = \sum_k \Omega(f_k)$，用 boosting 的方法进行优化。

$$\begin{aligned} Obj^{(t)} &\sum_{i=1}^{n} l(y_i, \hat{y}_i) = \sum_k \Omega(f_k) \\ &= \sum_{i=1}^{n} l(y_i, \hat{y}^{(t-1)} + f_t(x_i)) + \Omega(f_k) + constant \end{aligned} \tag{6}$$

从而找出最佳化的 f_t。

xgboost 是 Gradient Boosting 的一种高效系统实现，并非单一算法。xgboost 的基学习器既有树（gbtree）又有线性分类器（gbfinear），从而得到带 L1＋L2 惩罚的线性回归或者逻辑回归。xgboost 对损失函数采用了二阶泰勒展开，因而利用的是二阶导数信息。

Xgboost 将模型的复杂度作为正则项添加在目标函数来优化模型，并添加了对于后期的剪枝处理，提出 shrinkage 以及列（特征）二次采样的方法。

xgboost 是大规模并行 boosted tree 的工具，它是目前最快最好的开源 boosted tree 工具包，比常见的工具包快 10 倍以上。在数据科学方面，有大量 kaggle 选手选用它进行数据挖掘比赛，其中包括两个以上 kaggle 比赛的夺冠方案。在工业界规模方面，xgboost 的分布式版本有广泛的可移植性，支持在 YARN，MPI，Sungrid Engine 等各个平台上面运行，并且保留了单机并行版本的各种优化，使得它可以很好地解决于工业界规模的问题。

3.6 组合预测模型

组合预测模型实质是多分类器进行组合，其目的是为了将单个分类器（也叫基分类器 base classifier）进行组合，提升对未知样本的分类准确率。构建组合分类器的逻辑视图可以用以下的图表示：